Fuzzy C-Means

Andrei Paraschiv - Data Mining, Sem 1, 2018-2019 / Master ACS

## 1. Introduction

Fuzzy C-Means is an unsupervised clustering method developed by Dunn in 1973 [1] and improved by Bezdek in 1981 [2]. Clustering is one of the most important unsupervised learning problems - finding a structure in a collection of unlabeled datapoints.

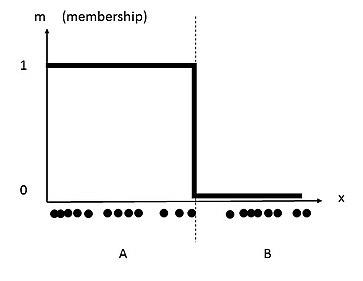
Clustering can be viewed as the process of organizing objects or events into groups of members that share a high number of similarities.

The goal of Fuzzy C-Means clustering is to determine the intrinsic grouping of a dataset by allowing a datapoint / object / event to have a degree of membership into each group. Any individual datapoint can belong to two or more clusters based on a membership function. In real applications very often there are no sharp borders between groups, this is why a fuzzy clustering method is often preferred to a hard clustering one and the results can identify better the internal structure of the data.

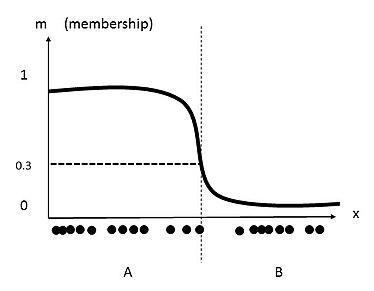
## 2. Algorithm description

Fuzzy C Means algorithm is a clustering method, an extension of K Means, that allows one data point to belong to two or more clusters based on a membership function. This clustering method is frequently used in pattern recognition.

In a hard clustering algorithm, the membership function the membership function takes the shape of a rectangular function



while in a fuzzy clustering method we might have a membership function similar to a sigma shaped function



The goal of the FCM algorithm is to minimize the objective function

where wij is the degree of membership for the datapoint xi to the centroid cj and m>=1 is a hyperparameter for the clustering which determines the fuzziness of the clusters.

The centroids are computed using :

and the membership degrees :

In order to start we need to choose a number of clusters and randomly assign data points to this clusters.

For each iteration we compute the centroids for each cluster, for each data point we compute the membership degree for each cluster based on these new centroids.

We repeat the iterations until the maximum change in membership degrees between two iterations is less then the sensitivity threshold ε

To summarize, the algorithm steps are:

1. Select an initial fuzzy partition by random (assign the values for all wij )
2. compute the centroids for each cluster using the fuzzy partition
3. update the fuzzy partition (the values for wij )
4. repeat step 2 and 3 until the centroids do not change or the change is below a specific threshold ε

## 2. Importance and practical applications

Clustering is an important classification method. FCM, being an unsupervised machine learning algorithm with an relative low order of complexity, it can be easily used for clustering large datasets with good results.

Due to the nature of fuzzy logic that deals with approximate and non-precise logic, FCM is more suited for coping with the nature of reality then hard clustering methods.

There are many usages for FCM in data analysis, pattern recognition, image segmentation.

Main areas of implementation are

* *Marketing*: customer segmentation, finding customers with similar behaviour or interests
* *Biology*: classification of plants, animals
* *WWW*: document classification, web log analysis
* *Insurance*: grouping customers by risk, finding unusual risk pattern, fraud detection
* *City Planning*: grouping houses by value, type and geographical location
* *Medicine:* patient classification, clustering in medical diagnostic systems

## 3. Pros and Cons

Some of the main advantages of Fuzzy C-Means are:

* The algorithm always converges
* It is unsupervised
* Leads to a more natural approach to clustering
* It has a relative simple implementation
* The implementation is relativ simple

The weak points and issues encountered are

* The complexity is higher than K-Means - O(ndc²i) vs O(ndci)
* It has a sensitivity to the initial guess
* Since all weights for a given point add up to 1 - it has a sensitivity to outliers and noise
* The effectiveness of the method is depending on the definition of the distance

## 4. Datasets used

### a) Weightlifting

The Powerlifting datasource is the

Nr Clusters: 5

Nr Samples: 131847

Cluster C1:

38.65

80.06

Cluster C2:

21.65

68.68

Cluster C3:

32.51

193.69

Cluster C4:

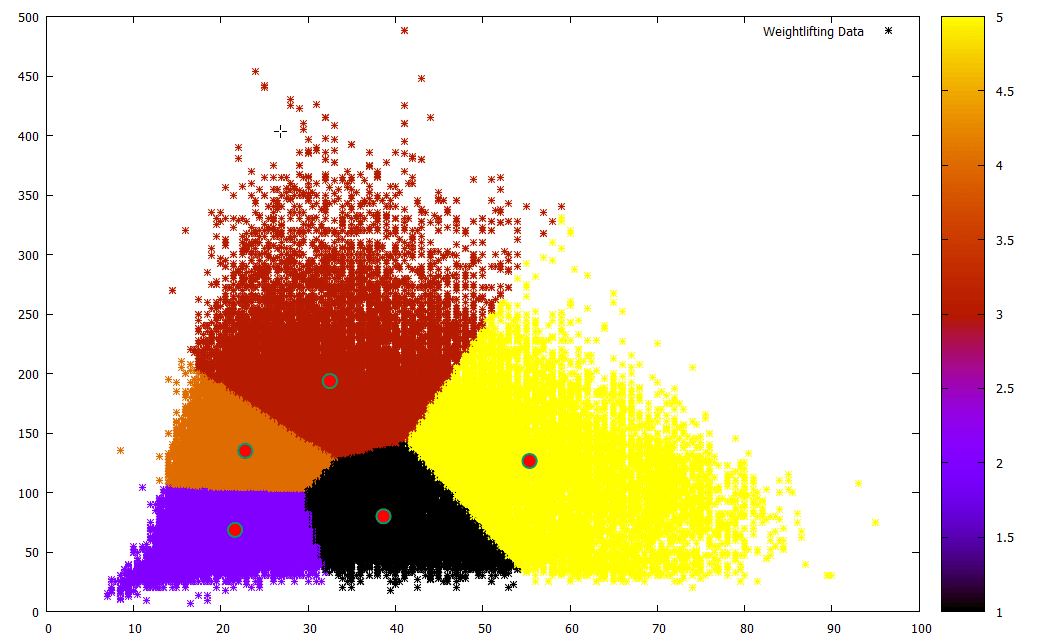
22.82

135.08

Cluster C5:

55.37

126.58



## 5. Results

## 6. References

[1] J. C. Dunn (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics, 3:3, 32-57, DOI: 10.1080/01969727308546046

[2] Bezdek, James. (1981). Pattern Recognition With Fuzzy Objective Function Algorithms. 10.1007/978-1-4757-0450-1.

[3] Wikipedia, https://en.wikipedia.org/wiki/Fuzzy\_clustering

https://www.win.tue.nl/~mbottoto/files/papers/conference\_papers/fuzzy\_citi2017.pdf